Capital Bike Share Predictive Model Report

Prepared By: Clark P. Necciai Jr.

November 09, 2023

Contents

1 Executive Summary

Capital Bike Share provides a network of multi-purpose bicycles to the denizens of the Washington D.C. Metropolitan region. We have been approached by Capital Bike Share to delve into their hour-by-hour observations over the 2011 and 2012 years Preliminary insights into the dataset provided to us from Capital Bike Share revealed that demand for usage of these bicycles can be affected by a variety of noteworthy, influential factors.

2 Problem Statement

We have been tasked with identifying the most influential variables relative to their predictive power affecting the amount of total hour-by-hour bicycle rentals. Beginning with an exploratory data analysis and inspection of our dataset, we aim to systematically determine these variables with respect to their predictive power through a well-formulated, multiple linear regression model.

Our modeling process will culminate in a model which will not only accurately predict rental demand, but simultaneously provide an assumption-verified, statistical performance evaluation confirming our models' reliability. Below we provide the methodology of our approach, beginning with an inspection and analysis of our data, followed by our modeling selection strategy and diagnostic testing to ensure model generalizability. Finally, we conclude with our recommendations and takeaways.

3 Methodology

3.1 Data Preprocessing

We began our approach by taking an overview of the integrity of our dataset. We discovered no missing values nor duplicated observations. We did, however, note variables which would be irrelevant to the overall analysis and took steps to filter out these irrelevant features.

Notes: Variables were renamed for descriptive clarity. No missing nor duplicate data detected. We reevaluated each variables underlying data type corresponding values, and when necessary, replaced each with more appropriate types and corresponding values. [Explain why you did this?] Make note of the changes that happened as a result and the final dimensionality of the dataset after pre-processing.

3.2 Exploratory Data Analysis

3.3 Univariate Analysis

3.3.1 Target Variables

Each of our three target variables could be seen to have vastly positively skewed distributions.

3.3.2 Predictor Variables

3.3.2.1 Categorical Variables

3.3.2.1.1 `date` An inspection of the `date` variable's distribution revealed that not every discrete value of date has an expected equivalent number of hourly interval measurements. For the **731** distinct date values present in our data set, when grouped by date, the majority of each discrete date contained twenty-four or twenty-three of the expected hourly observations. However, fourteen of the dates contained less and in some cases contained a mere single hourly observation, which in total equated to **233** hours worth of missing observations. While the fact that these distinct dates have missing hour-to-hour observations, it does not immediately follow that these observations are necessarily outliers. However, it should be noted that Capital Bike Shares' hour-to-hour observational system has hourly gaps, which would otherwise provide useful descriptive and predictive analytics.

3.3.2.1.2 `season` The seasonal distribution with regard to the number of recorded observations between Fall, Winter, Spring, and Summer revealed only a slight imbalance. Fall contained the fewest of these with 4232 observations, Winter with 4242 observations, Spring with 4409 observations, and Summer contained the most with 4496 observations. As a whole, the distribution was evenly distributed with each season containing approximately within +/- 1% of a quarter of the observations as would be expected.

3.3.2.1.3 `year` The distribution of observations for the two recorded years, 2011 and 2012, were nearly exact at 49.74% and 50.26%, respectively.

3.3.2.1.4 `month` Across all monthly grouped observations, the distribution of the `month` variable was approximately uniform. However, the month of February appeared unique with it having the fewest number of observations of 1341. May had the highest count of recorded observations with 1488 recorded. All other months' counts of observations fell within these two counts of 1341 to 1488 (about 8% each of total observations).

3.3.2.1.5 `hour` When observing a sorted hourly distribution from 12:00AM to 11:00PM, we found that the count of each distinct hour were nearly equal. However, a trend in the number of observations can be seen with a decrease beginning at approximately 1:00AM and continuing until 3:00AM when it then increases through 6:00AM.

3.3.2.1.6 `day` The value counts for the `day` variable were all nearly approximate.

3.3.2.1.7 `weather` We found an immediately noticeable yet unsurprising distribution with the `weather` variable. **11413** of the observations were of Type 1, indicating the majority of our observations were recorded in favorable weather conditions. **4544** observations were of Type 2 and **1419** were of Type 3. A mere **3** observations were recorded for Type 4.

3.3.2.2 Continuous Variables

3.3.2.2.1 `temp`

3.3.2.2.2 `atemp`

3.3.2.2.3 `hum`

3.3.2.2.4 `windspeed`

3.3.2.3 Boolean(True/False) Variables

3.3.2.3.1 `holiday` Inspection of our `holiday` variable revealed dates which should have been marked as holidays and others which should not have been. We decided to re-label these observations based on official federally recognized holidays.

In comparison to our dataset: **2011-01-01(New Years), 2011-12-25(Christmas), 2012-01-01(New Years), and 2012-11-11(Veterans Day)** were incorrectly mislabeled as *not* being holidays. **2011-12-26, 2012-01-02, and 2012-11-12** were mislabeled as **being** holidays. The aforementioned dates were corrected as to maintain consistency with federally recognized holidays. **2011-04-15 and 2012-04-16**, while not being federally recognized holidays, are dates of observance for Emancipation Day in the Washington D.C. Area. Due to the similar holiday-like observance of Emancipation Day in our area of interest, these two dates will be labeled as holidays.

The final distribution remains highly disproportionate yet expected, with only 525 of 17379observations being labeled as holidays. The remaining **16854** were not holidays.

3.3.2.3.2 `workingday` The distribution reveals that the percentage of total observations being labeled as a `workingday` are nearly twice the amount as not being labeled.

3.4 Correlation

We decided to investigate some of the stronger, more notable relationships made apparent by the correlation matrix just prior.

Positive correlations between our temperature related variables and target. hum negatively related to our target variables

3.5 Notable Multivariate Analysis

3.5.0.1 Temperature Related Variables Examining the multi-variable relationship between our temperature related variables, `temp`/`atemp`, and the target variables revealed the presence of multi-variate outliers.

These twenty-four observations had the exact same `atemp` value of **0.2424**, but `temp` values which were relatively high and variable in relation, ranging from **0.62 to 0.86**. Delving into these observations values, we discovered that they were all recorded sequentially, on the exact same day(*2012-08-17*) and under the same weather conditions(*Type 1*). The changing levels of humidity and wind speed lead us to further believe that the real feel temperature should have likewise varied when compared to other observations of similar values in the dataset.

When considering the evidence of these twenty-four `temp` and `atemp` multivariate as being observations, we will opt to remove these observations. This removal helps to ensure data integrity as we move towards the modeling process.

3.5.0.2 Hourly Usage and Day Due to the widespread availability of bicycles as a mode of transportation at all hours of the day, our intuition led us to investigate possible patterns of usage based solely on time. What we uncovered, showed that during working days, for target variables `registered` and `count_rentals`, there is strong indication that bicycles are being utilized as a primary mode of transportation to and from work. No strong pattern(constant trend) of usage was seen for `casual` rentals.

For the `registered` and `count_rentals`, there is a notable pattern in bicycle rentals during commute hours to/from work. It may be the case that Capital Bike Share bicycles are used as a primary mode of transportation for the majority of riders during working hours. Before 9:00AM, we note an upward trend of rentals followed by a stark decrease just after. We are interpreting this as individuals arriving at work. This is followed by high usage at 5:00PM when people typically leave work, where the bikes may be rapidly available. This is then followed with slowly decreasing

usage. Based on these insights, we will be considering including an interaction term between `hour` and `workingday`(or `day`)in our models predicting `registered` and `count_rentals` to capture this variability in rental numbers. Doing so will also allow us to capture the high usage trends during weekend hours as well.

3.6 Feature Engineering

3.6.0.1 Target Variables

3.6.0.2 Predictor Variables

3.6.0.2.1 `date` We speculated that the creation of **731** distinct variables representing `date` could have severe impacts on model performance when regressing our target variable. The only unique piece of information provided by `date` is the day of the month. All other information is already found in our `year` and `month` variable.

Our team experimented with the inclusion of a variable representing this day of the month(`day_of_month`) in an attempt to achieve better overall model performance. After extensive testing, we found that *exclusion* of this variable and of `date` yielded better performance.

While inclusion of the variables helped explain some error(RSS), the trade-off of having to create additional variables to accommodate the model fitting process ultimately proved detrimental to overall performance.

It is for this reason, our team has decided to remove the `date` variable from the modeling process entirely.

3.6.0.2.2 Dummy Variable Creation

4 Model Building

4.1 Data Partition

4.2 Initial Full Model(s)

4.2.1 Regression Assumptions

4.2.1.1 Normality

4.2.1.2 Homoscedasticity

4.2.1.3 Multicollinearity

4.2.2 Model Improvement

4.2.2.1 Model Outliers/Influential Observations

4.2.2.2 Target Variable Transformation [REWRITE]

during this analysis is to fit multiple linear regression models that carry certain assumptions which must be satisfied. These mainly include our response variable being linearly related to the predictors(linearity) and equality of variances in the residuals(homoscedasticity). With their current skewed distributions, these two assumptions would undoubtedly be violated post-model building.

In order to satisfy the assumptions of linear regression, a transformation on each of the target variables becomes necessary. We considered multiple transformations to approach normality, including, the Box-cox, logarithmic, square-root, cube-root, and fourth-root transformations. For each of our target variables, the transformation which approximated normality (closest skewness value to 0) the closest was the **cube-root** transformation. Neither the logarithmic and Box-cox transformations could not be applied to `casual` and `registered` due to the presence of zero(0) values which are incompatible with the transformations.

4.2.2.3 Multicollinearity (VIF)

4.2.3 Reduced Models (Feature Selection)

5 Conclusions and Recommendations

6 Appendix